Annotators' Certainty and Disagreements in Coreference and Bridging Annotation in Prague Dependency Treebank

Anna Nedoluzhko
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
nedoluzko@ufal.mff.cuni.cz

Jiří Mírovský
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
mirovsky@ufal.mff.cuni.cz

Abstract

In this paper, we present the results of the parallel Czech coreference and bridging annotation in the Prague Dependency Treebank 2.0. The annotation is carried out on dependency trees (on the tectogrammatical layer). We describe the inter-annotator agreement measurement, classify and analyse the most common types of annotators' disagreement. On two selected long texts, we asked the annotators to mark the degree of certainty they have in some most problematic points; we compare the results to the inter-annotator agreement measurement.

1 Introduction

The coreference and bridging annotation in the Prague Dependency Treebank (PDT) is one of the largest existing manually annotated corpora for pronominal, zero and nominal coreference and bridging relations. Contrary to the majority of similarly aimed corpus projects (Poesio 2004, Poesio – Artstein 2008, Poesio et al. 2004, Recasens 2009, Krasavina – Chiarchos 2007, etc.), coreference and bridging relations have been annotated directly on the syntactic trees and technically they are a part of the tectogrammatical (complex semantic) layer of PDT. This approach allows us to include relevant syntactic phenomena annotated earlier (such as e.g. appositions, coreference relations between subject and predicate nominals, etc.) into the coreference representation, and to take advantage of the syntactic structure itself (resolution of elliptical structures, coordinations, parentheses, foreign expressions and identification structures, direct speech, etc.)[1]. Also, from the perspective of querying and visualizing the treebank, all the different types of linguistic information are interlinked, available and visible at once. One of the important advantages is that PDT includes information on topic-focus articulation (Hajič et al. 2006) and discourse annotation (Mladová 2011).

Comparing the results of inter-annotator agreement in manual annotations of language phenomena at different language levels makes evident that the degree of agreement goes down when proceeding from phonological to "higher" language levels. On the one hand, relations that cross the sentence boundary are not so systematically described both in classical linguistics and in annotation guidelines, causing disagreements due to different understanding of terms. On the other hand, such relations are much more vague and in many cases ambiguous. Both these problems influence the measurement of the inter-annotator agreement. In this paper, we present results of the inter-annotator agreement measurement for nominal coreference and bridging relations for Czech and compare them to the degree of certainty the annotators had while marking these relations.

2 The Annotation Scheme

Within the bounds of coreference-like phenomena, three types of relations are marked in PDT:

a) grammatical coreference (coreference of relative and reflexive pronouns, verbs of

[1] The benefits of the tectogrammatical structure for coreference annotation are described in detail in Nedoluzhko – Mírovský (2013).

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 236–243,
Prague, August 27–30, 2013.

control arguments, arguments in constructions with reciprocity and verbal complements),

b) pronominal and nominal textual coreference (including zero anaphora), which is further specified into coreference of specific (type SPEC) and generic (type GEN) noun phrases, and

c) bridging relations, which mark some semantic relations between non-coreferential entities.

The following types of bridging relations are distinguished: PART-OF (e.g. *room - ceiling*), SUBSET (*students - some students*) and FUNCT (*state - president*) traditional relations (see e.g. Clark 1977), CONTRAST for coherence relevant discourse opposites (*this year - last year*), ANAF for explicitly anaphoric relations without coreference, e.g. for metalinguistic references (*rainbow - that word*) and the further underspecified group REST[2].

Grammatical coreference typically occurs within a single sentence, the antecedent being able to be derived on the basis of grammar rules of a given language. For this reason, grammatical coreference is the least ambiguous among the coreference types, its annotation is the most reliable, being close to other grammatical phenomena annotated in PDT.

3 Solving Coreference Ambiguity in Similar Projects

Problems of low inter-annotator agreement and ambiguity in annotation of coreference and bridging relations have been topics of active discussions during the last few years. Shortcomings of straightforward definitions of coreference were pointed out in Poesio and Artstein (2005). They were later analyzed in detail using linguistic and computational methods in Versley (2008), and Recasens et al. (2010, 2011). The group of so called "near-identity" relations, where the discourse entities to which the noun phrases refer cannot be called coreferential in all senses but still are rather coreferential than not, was separated from the cases of full-coreference. Coreference was thus redefined as a scalar relation between linguistic expressions that refer to discourse entities considered to be at the same granularity level relevant to the linguistic and pragmatic context (Recasens et al. 2011). The "near-identity" relation holds e.g. between *several hundred disabled people* and *the congregated* in Versley's (2008) example (1). The groups of people addressed by these noun phrases are not the same but the difference is neutralized by the context:

> (1) *For a "barrier-free Bremen," several hundred disabled people went onto the streets yesterday—and demonstrated for "Equality, not Barriers." . . . "Why always us" the congregated asked on the posters*.

However, the attempt to annotate "near-identity" explicitly has proved to be unreliable, because it is difficult for annotators to recognize such relations (Recasens et al. 2012). Also ambiguity seems to be much better identified not by asking annotators to code ambiguous expressions but by comparing the annotations produced by different annotators (Poesio and Artstein 2005). Explicitly marked ambiguity is annotated in the PoCoS corpus for German (Krasavina – Chiarchos 2007) but was not analysed in detail yet.

4 Evaluation of Parallel Annotations

F-1 on textual pronominal coreference (including zeros)	0.86[3]
F-1 on textual coreference for specific NPs	0.705
F-1 on textual coreference for generic NPs	0.492
F-1 on bridging relations	0.455
new textual kappa of agreement on type	0.759
bridging kappa of agreement on type	0.889

Table 1: Evaluation of parallel annotations

In order to evaluate the inter-annotator agreement on selected texts annotated by two or more annotators, we used F_1-measure for the agreement on arrows and Cohen's κ (Cohen 1960) for the agreement on types of arrows. During the annotation period, 11 measurements between two coders have been

[2] For a detailed classification of identity coreference and bridging anaphora used in PDT, see e.g. Nedoluzhko - Mírovský (2011).

[3] As reported in a technical report from the annotation of PDT (Kučová et al. 2003).

provided for (in total) 1,606 sentences in 39 documents.

Table 1 shows average results of the inter-annotator agreement measurements for all types of textual coreference and bridging relations.

5 Cases of Typical Disagreement

Proceeding to further phases of the annotation process didn't give us any dramatic enhancement of the inter-annotator agreement. Some later measurements have shown even lower agreement than the earlier ones, although the quality of annotating was very high. That indicated that the results primarily depend neither on the annotators' experience in the field nor on their ability to follow the guidelines.

Technically, as for the annotators, four general issues appeared to be difficult to decide: whether the relation is to be annotated for coreference/bridging at all, what is the correct antecedent of a given noun phrase, to distinguish between the bridging anaphora and the textual coreference and to select the type of the bridging anaphora or the textual coreference. These issues are closely analysed in the sections 5.1 to 5.4, with real-data examples.

5.1 Annotating / not annotating a relation

There is a relatively high degree of disagreement in the very recognition of a coreference or bridging relation in some typical cases.

The most frequent example is a general reference of noun phrases, which may and may not be annotated as coreferential.

(2) *A když už byla knížka hotova, tak se zjistilo, že je praktická i pro rodiče. V této knize je poučení, jak snášejí děti_rozvod a jak na něj reagují, a návod, jak se mají rodiče chovat, aby se utrpení dětí snížilo. (=After the book had been already written, it was clear, that it is quite useful for parents too. The book contains explanations, how children go through divorce, how they react to it, and the instructions how parents should behave to minimize the suffering of their children.*.)

The disagreement is even more likely if the generic antecedent is relatively far from the noun phrase in question (example 3):

(3) *Preferuji širší předvedení s mnoha vnitřními souvislostmi, protože nám chybějí kritéria pro hodnocení současné české výtvarné kultury*. {11 sentences inbetween} *Měli bychom se znovu pokusit ... získávat současné umění, abychom jednou měli autentický soubor naší doby* (= *I prefer wider demonstration with many internal connections because we lack criteria for evaluation of contemporary Czech art. We should try ... to acquire the contemporary art again, in order to get an authentic set of our time.*)

5.2 Different selecting the antecedent / anaphoric element

Compare (4) - (6) for identity coreference. In (4), the anaphoric noun phrase *the new structure* corefers with *the type F railing* in one coder's annotation and with *the G Street Bridge* in the other's.

(4) *In Richmond, Ind., the type F railing is being used to replace arched openings on the G Street Bridge. Garret Boone, who teaches art at Earlham College, calls the new structure ``just an ugly bridge'' and one that blocks the view of a new park below.*

The measure in (5) corefers with *the House bill on airline leveraged buy-outs* in one coder's annotation and with the extended noun phrase *legislation similar to the House bill on airline leveraged buy-outs* in the other annotation:

(5) *While the Senate Commerce Committee has approved legislation similar to the House bill on airline leveraged buy-outs, the measure hasn't yet come to the full floor.*

The following example (6) demonstrates disagreement in constructions with measure and time-period words. *The year earlier* may corefer with *prior-year* or *the prior-year period*:

(6) *That compares with operating earnings of $132.9 million, or 49 cents a share, the year earlier. The prior-year period includes...*

In (7), *Tajikistan* is linked by the bridging SUBSET relation by both coders but they chose different antecedents: one coder linked it to *these countries* (thus coreferring it to the whole coordinating construction *post-*

communist countries of Eastern Europe and the republics of the former USSR.), while the other coder made a more precise decision and linked *Tajikistan* to *the republics of the former USSR*, i.e. just to one element of the coordination. Both annotations are empirically correct, the decision depends on the coder's world knowledge.

> (7) *Tiskárny bankovek mají i nové zákazníky, především v postkomunistických zemích východní Evropy a republikách bývalého SSSR. Bankovky v těchto zemích jsou náchylné na padělání a mají zastaralý design. Kanadská firma CBNC bude tisknout nové bankovky pro Tádžikistán* (= *They have new clients, first of all in the post-soviet countries of East Europe and in the republics of the former USSR. Banknotes in these countries can be easily falsified. The CBNC Company will print banknotes for Tajikistan.*)

5.3 Distinguishing between the bridging relations and the textual coreference

Disagreement in choosing between bridging relations and identity coreference relations are often the case when noun phrases have a generic or unspecific reference. In (8), the relation between *banknotes* and *undamaged banknotes* is understood as coreference by one coder and as bridging SUBSET by the other one.

> (8) *I přes klesající inflaci ve světě ... je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví.* […] *... Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek*. (= *Although inflation in the world rather decreases, … printing banknotes and production of banknote paper is still one of the most profitable areas. Mass expansion of ATMs calls for permanent increase of undamaged banknotes*.)

5.4 Selecting the type of the bridging anaphora or the textual coreference

As for bridging relations, some relations can be disagreed on in different contexts, e.g. the relation between *Slovakia* and *Bratislava* in (9) may be understood from two different points of view. One coder marked it geographically (Bratislava is a part of Slovakia – relation PART-OF), the other understood the relation from the point of view of its function (Bratislava is the capital of Slovakia – relation FUNCT).

> (9) *Slovensko po několika měsících diskusí devalvovalo svou měnu o deset procent. […] Spíše je otázkou, zda Bratislava nepřistoupila k akci poněkud pozdě. (=After several months of discussions, Slovakia devalued its currency by ten percent. […] The question is whether Bratislava was not somewhat late with this decision.)*

For identity coreference, types SPEC and GEN were distinguished (see section 2) according to which noun phrases the coreference relation was applied to. However, in real corpus examples, the distinction is not always clear and coders may mark it differently.

6 Reasons for Disagreement

The evaluations of parallel annotations of selected texts brought up some interesting observations. The nature of disagreements corresponds to the general problem of a formal description on such a high level of language, namely – the texts sometimes allow for different, equally relevant interpretations. Moreover, the guidelines restrict us by the number of arrows leading from one node, and only a few formalized types of coreference and bridging relations are annotated in PDT, thus it does not fully reflect the real situation of text cohesion. See e.g. (4), where the semantically correct decision would be to annotate both relations as (near-)coreference, but not disposing such rich annotation guidelines, coders have to choose one variant and disagreement is to be expected.

Reflecting the results, we were able to distinguish two main textual factors for disagreement: the text size and the degree of its abstractedness. Especially long texts with a large number of generic nouns, abstracts and deverbatives have the lowest inter-annotator agreement.

A detailed manual comparison of parallel annotations revealed that almost three quarters of the coders' disagreements come from the text ambiguity (the relations may be empirically ambiguous as in (5), where coreferring with different antecedents may change the meaning, or rather near-identical in the sence of Recasens (2010), when different interpretations are possible that do not actually change the meaning of the text as a whole).

Constructions with nouns of measure and time periods appear to be hard to agree on (see e.g. 6) – in spite of quite detailed descriptions in the guidelines, coders tend to mark them differently in different types of context according to their intuition in every particular case. Generic noun phrases, abstract nouns and deverbatives cause really rich ambiguity in almost all coreference annotation projects. However, for Czech, it results in even more disagreements (examples (2), (3) and (8)), because Czech does not have grammatical means to mark definiteness, thus forcing not to make any distinction in marking coreference between definite and indefinite noun phrases.

Marking coreference between indefinite noun phrases results in a further reason of disagreement, and that is a different level of thoroughness of the coders' interpretation. For example, in (3), the antecedent for *the contemporary art* was used 11 sentences before, the noun phrase in question is positioned as new (it has focus value in the TFA-annotation) and a coder doesn't need to see any serious reason to connect it by a coreference relation with such a distant antecedent. The similar situation is in (10) where, although not distant, the identity of *the safety and health deficiencies* and *the hazards* is up to the coder's intuition.

(10) *Gerard Scannell, the head of OSHA, said USX managers have known about many of the safety and health deficiencies at the plants for years, ``yet have failed to take necessary action to counteract the hazards."*

The rest of the coders' disagreements are caused be either a coder's mistake (cca 15% of occurences) or guidelines inconsistency (cca 10% of occurences).

7 Certainty of the Manual Annotations

To find out which part of problematic cases the coders are aware of, we organized one special inter-annotator agreement measurement. We asked the annotators to annotate the data as usual and also mark the certainty they had in several parts of the task.[4]

They were asked to mark the certainty for their annotation decisions on the scale of 1 to 3 (1 means quite certain, 2 means moderately certain, 3 means not really certain). The certainty was marked for four types of decision (tasks), according to cases of frequent disagreement described in sections 5.1-5.4, i.e. certainty in the presence of a relation, certainty in selecting the antecedent, certainty in distinguishing between the bridging relation and the textual coreference and certainty in selecting the type of the bridging anaphora or the textual coreference.

The certainty and the inter-annotator agreement were then measured separately for these tasks and (where applicable) also separately for various levels of certainty.

7.1 Certainty in the presence of a relation

Table 2 shows the average certainty the annotators expressed in various situations in the task of detecting the presence of a relation.

measurement	**average certainty**
one annotator marked a relation (bridging), the other has not marked any	1.88
one annotator marked a relation (coref-text), the other has not marked any	1.44
one annotator marked a relation (any relation), the other has not marked any	1.68
both annotators marked a relation (bridging)	1.35
both annotators marked a relation (coref-text)	1.17
both annotators marked a relation (any relation)	1.25

Table 2: Average certainty in the task of detecting the presence of a relation

The numbers show that the lower the agreement is, the less sure the annotators are. However, if we look at the absolute numbers of (non-)annotating textual coreference, we see that the number of cases where the annotators didn't mark uncertainty but still disagreed exceeds all other cases. In the analysed documents, uncertainty was marked in 26 cases of disagreement. In another 30 cases where only one coder annotated a coreference relation, the uncertainty was not marked.

7.2 Certainty in selecting the antecedent

Table 3 shows the inter-annotator agreement in the task of choosing the antecedent of the

[4] This measurement was performed on 190 sentences in 2 documents.

relations, depending on the relation and the certainty declared by the annotators. It is measured on the cases where both the annotators marked a relation at the given position in the data.

measurement	certainty declared by the annotators	agreement
bridging relations	both 1	48%
coref-text relations	both 1	67%
any relation	both 1	62%
bridging relations	at least one of them 2 or 3	33%
coref-text relations	at least one of them 2 or 3	36%
any relation	at least one of them 2 or 3	41%

Table 3: The inter-annotator agreement in the task of choosing the antecedent

Again, the numbers show a lower agreement in cases where the annotators were not sure about the antecedent. However, from 27 disagreements in choosing the antecedent, only 16 were marked as uncertain by at least one annotator.

7.3 Certainty in distinguishing between the bridging anaphora and the textual coreference

Table 4 shows the inter-annotator agreement in the decision whether the relation is a bridging anaphora or a textual coreference, depending on the certainty declared by the annotators. It is measured on the cases where both the annotators marked a relation at the given position in the data.

measurement	certainty declared by the annotators	agreement
any relation	both 1	97%
any relation	at least one of them 2 or 3	84%

Table 4: The inter-annotator agreement in the decision whether the relation is the bridging anaphora or the textual coreference

The difference in agreement between "certain" and "uncertain" relations in this case is not so relevant. As seen from the table, the agreement is very high. In most cases (21 out of 32), the annotators marked ambiguity but still made the same decision.

7.4 Certainty in selecting the type of the bridging anaphora or the textual coreference

The following table shows the inter-annotator agreement in the task of choosing the type of the bridging anaphora or the textual coreference, depending on the relation and the certainty declared by the annotators. It is measured on the cases where both the annotators marked a relation at the given position in the data.

measurement	certainty declared by the annotators	agreement
bridging relations	both 1	97%
coref-text relations	both 1	96%
any relation	both 1	92%
bridging relations	at least one of them 2 or 3	75%
coref-text relations	at least one of them 2 or 3	73%
any relation	at least one of them 2 or 3	63%

Table 5: The inter-annotator agreement in the task of choosing the type of the bridging anaphora or the textual coreference

8 Discussion

Analyzing the inter-annotator agreement together with the results of annotators' certainty about the relations reveals the following challenging issues:

Firstly, it points out the complexity of real corpus data which can never be reflected by any annotation guidelines in full detail. See e.g. examples (2), (4) and (6) that are not empirically ambiguous but cannot be captured by single yes/no identity rules. The same is true for bridging anaphora: a small set of relations which can yet be reasonable in large-scale corpora annotation cannot capture all cases of text cohesion. Unlike syntax, annotation of "higher" levels (coreference, bridging relations, discourse, etc.) does not reflect a language phenomenon as a whole. It rather excerpts a part of it, which is relevant for a certain task, and formalizes it to a reasonable degree. Contra-intuitivity, such formalized decisions result in a lower inter-annotator agreement. Also the annotators' certainty is lower in cases where intuition goes against the guidelines. Entities might seem to be very coherent, but there may be no good formal relation to be identified.

Secondly, empirical ambiguity seems to be more frequent on text level than on syntax level and lower. However, a detailed analysis of our data confirms the Recasens' at al. (2010) and Poesio-Artstein's (2005) statements: ambiguity is much better seen when comparing parallel annotations than when asking annotators to mark it by themselves.

Thirdly, weak points of the annotation guidelines are revealed. Not having precise and exhaustive rules, annotators naturally doubt more. In our case, this concerns first of all classifying generic noun phrases, abstract nouns and deverbatives. Also noun phrases with measures of different kind, time periods and some language specific constructions appear to be problematic. Annotators are much less certain about relations between generic and abstract nouns. Also the inter-annotator agreement for these cases is always lower than that for specific nouns with concrete meaning. Generally, we can say that in Czech, the most frequent reason for inter-annotator disagreement is not so much metonymy and different cases of near-identity relations in the sense of Recasens, but rather the relations between noun phrases with a generic and an abstract meaning. An improvement of such a problematic area would be to have the semantic information assigned to nouns themselves, as a part of tectogrammatical information. However, this task is very time-consuming.

Comparing the parallel annotations also shows that annotators are more sure about relations between noun phrases in topic and contrastive topic than about those in focus. More than other nouns, this fact concerns generic and abstract nouns and deverbatives. Coreference of these types of nouns in focus is not always obvious. Presented as new, coreference relation with a preceding noun phrase referring to the same type loses its relevance. However, this statement is rather a hypothesis, it needs further investigation.

9 Conclusion

We presented an evaluation and analysis of disagreements in the annotation of coreference and bridging relations in the Prague Dependency Treebank. As demonstrated by the results of parallel annotations, the agreement decreases in the direction from pronominal and zero coreference towards bridging relations. We extracted four most frequent types of problematic cases, exemplified them and described the possible reasons of inter-annotator disagreements. Then we asked annotators to mark the certainty they had in these cases and compared the results to the results of inter-annotator agreement. Although the percentage numbers were quite predictable (the less sure the annotators were, the lower was the agreement), the absolute numbers indicate that there remain many disagreements where uncertainty was not marked by any annotator.

Acknowledgments

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875).

References

Herbert Clark. 1977. Bridging. In Johnson-Laird and Wason, editors, *Thinking: Readings in Cognitive Science.* Cambridge, pp. 411–420.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37–46.

Jan Hajič et al. 2006. *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Olga Krasavina and Christian Chiarcos. 2007. PoCoS – Potsdam Coreference Scheme. *Proc. of ACL 2007*, Prague, Czech Republic.

Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo. 2003. *Anotování koreference v Pražském závislostním korpusu.* ÚFAL Technical Report TR-2003-19.

Lucie Mladová. 2011. *Annotating Discourse in Prague Dependency Treebank.* A presentation at the workshop Annotation of Discourse Relations in Large Corpora at the conference Corpus Linguistics 2011 (CL 2011), Birmingham, Great Britain, July 2011.

Anna Nedoluzhko and Jiří Mírovský. 2013. How dependency trees and tectogrammatics help annotating coreference and bridging relations in Prague Dependency Treebank. *Proceedings of the International Conference on Dependency Linguistics* (*Depling 2013*), Prague, Czech Republic (in press).

Anna Nedoluzhko and Jiří Mírovský. 2011. *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank.* Annotation manual. Technical report No. 44, ÚFAL, Charles University in Prague.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II*. Ann Arbor, pp. 76–83.

Massimo Poesio, Rodolfo Delmonte, Antonella Bristot, Luminita Chiran, Sara Tonelli. 2004. *The Venex corpus of anaphora and deixis in spoken and written Italian*. Manuscript.

Massimo Poesio. 2004. The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. *Proceedings of The 5th SIGdial Workshop on Discourse and Dialogue,* Boston.

Massimo Poesio, Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. *Proceedings of LREC 2008*, Marrakech.

Marta Recasens, M. Antònia Martí. 2009. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation.*

Marta Recasens, Eduard Hovy, M. Antònia Martí. 2010. *A typology of near-identity relations for coreference (NIDENT*). In Proceedings of LREC 2010, Valletta.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, Non-Identity, and Near-Identity: *Addressing the complexity of coreference*. Lingua, 121(6), pp. 1138–1152 2011.

Marta Recasens, M. Antònia Martí, Constantin Orasan. 2012. Annotating Near-Identity from Coreference Disagreements. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul.

Yanick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation 6*, pp. 333–353.